양적형질유전자의 연관성분석을 위한 가계자료의 정보충분성 연구

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

이 수 옥

양적형질유전자의 연관성분석을 위한 가계자료의 정보충분성 연구

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2004년 12월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
이 수 옥

감사의 글

많은 고민과 함께 시작했던 대학원 생활이 벌써 마무리를 준비해야 할 시점이라고 생각하니 느낌이 새롭습니다. 2년이란 시간을 돌아보면 어려움들도 있었지만 주위의 많은 도움으로 더 큰 배움을 얻어갈 수 있었다고 생각됩니다.

대학원 생활 내내 지속적인 가르침을 주신 김동기 선생님께 감사드립니다. 유전통계라는 새로운 분야를 알아갈 수 있도록 도와주신 임길섭 선생님과 바쁘신 중에도 논문에 많은 관심과 조언을 아끼지 않으셨던 이학배 선생님, 통계의 적용을 고민하게 해 주신 조진남 선생님께 감사의 말씀을 전합니다.

누구보다 지금의 절 가장 많이 격려해주는 사랑하는 가족에게 감사드립니다. 제가 가는 길을 누구보다 응원해 주실 아버지께 감사함을 전하며, 곁에 계시지 않아도 항상 가장 큰 힘이 되고 있다는 것을, 가장 먼저 떠오르는 분이라는 것을 아실 것으로 믿습니다. 진학으로 고민할 때 누구보다도 힘을 실어 주셨던 엄마께 감사드립니다. 늘 믿어주시고, 힘겨운 일들에도 든든한 울타리가 되어 주시는 엄마께 이젠 보답드릴 날을 약속드리고 싶습니다. 언니를 대신해 큰 짐을 지고 나가는 동생 수진에게도 감사하다는 말을 전합니다. 늘 원하는 일을 할 수 있길 바라며 좋은 결실로 보답받길 진심으로 기원합니다. 어느 순간 훌쩍 커버린 듯한, 그래서 많은 의지가 되는 막내 창후에게도 감사의 말을 전합니다. 같이 하는 시간은 짧지만 늘 마음으로 응원하고 있다고 전하고 싶습니다.

논문이 완성되기까지 많은 조언을 해주신 송기준 박사님께 감사드립니다. 날카로운 지적들이 더 나은 모습을 위한 애정 어린 조언이라는 것을 알고 있습니다. 같이 고민해 주고 많은 도움을 주신 성민오빠, 무영오빠, 미영씨께 감사드립니다. 처음 대학원으로의 끈을 쥐어주고 대학원 생활 내내 따뜻한 위로와 충고를 해준 찬미언니께 감사를 전합니다. 마지막 학기를 맞는 원열오빠와 혜리씨는 좋은 결실을 맺기 바라며, 소연, 민진, 성은도 앞으로 남은 대학원 생활을 멋지게 마무리 할 수 있게 되길 바랍니다.

8명의 동기로 시작해 마지막까지 함께 한 신영, 은혜에게 고맙다는 말을 전합니다. 서로를 격려하고 위로하며 보낸 2년을 돌아보면, 혼자서는 해내지 못했을 것이라 느껴집니다. 계속 학업을 이어갈 신영과 새롭게 사회로 나가려는 은혜 모두에게 항상 좋은 일이 함께 하길 진심으로 기원합니다.

대학동기 지혜, 영란, 민경에게도 감사의 말을 전합니다. 서로 다른 길을 가는 중에도 항상 서로에게 힘이 되고 의지되었던 것처럼 앞으로도 함께 할 수 있길 바랍니다. 어느새 10년지기가 되어버린 선희, 원경, 민성, 혜진, 선미, 지웅, 순만, 재학, 선미등 소리모아 동기들에게 고마움을 느낍니다. 함께 대학생활을 시작할 때부터, 각각의 다른 삶을 위해 열심히 사는 오늘까지 늘 내편에서 기쁜일, 슬픈일들을 내일처럼 생각해준 친구들에게 감사를 전하며 나도 그들에게 위로가 되는 존재였기를 바랍니다.

마지막으로 늘 곁에서 격려와 위로를 아끼지 않고 긴 시간 동안 한결같은 마음으로 응원해주고 따뜻하게 감싸주었던 병희 오빠께 사랑과 감사의 마음을 전합니다. 앞으로 다가올 시간들도 함께 사랑하며 서로에게 버팀목이 되어 줄 수 있기를 희망합니다.

2004년 12월

이 수 옥 올림

차 례

그림차례 ····· iii

표차례 ····· iii

국문요약 ····· iv

Ⅰ. 서 론 ····· 1

1.1 연구배경 ····· 1

1.2 연구목적 및 방법 ····· 2

Ⅱ. 정보충분성 ····· 4

2.1 개요 ····· 4

2.2 일반가계자료의 질적형질 정보충분성 ····· 4

2.3 형제자료의 양적형질 정보충분성 ····· 6

Ⅲ. 일반가계자료의 양적형질 정보충분성 ····· 9

3.1 개요 ····· 9

3.2 유전모형 ····· 9

3.3 정보충분성 지표 ····· 11

3.3.1 가계의 유전적 구성 ····· 12

3.3.2 공분산 행렬 ····· 14

3.3.3 우도함수 ····· 15

3.3.4 가중치 ····· 17

3.3.5 최종 정보충분성 지표 ····· 18

Ⅳ. 실제자료를 이용한 정보충분성 측정 ····· 19

4.1 자료와 방법 ····· 19

4.1.1 표식유전자 및 양적형질의 선택 ····· 19

4.1.2 자료의 형태 ····· 20

4.2 분석방법 ····· 23

4.3 분석결과 ····· 25

4.3.1 정보충분성을 이용한 가계 선별 ····· 32
4.3.2 전체자료의 정보충분성 측정 ····· 32
Ⅴ. 결론 및 고찰 ····· 35
참고문헌 ····· 37
ABSTRACT ····· 41

그 림 차 례

그림 1. 4인 가계에서 표식유전자 정보유무에 따른 HDL의 정보충분성 지표 ····· 26
그림 2. 4인 가계에서 표식유전자 정보유무에 따른 HDL의 정보충분성 순위 ····· 26
그림 3. 표식유전자 정보유무에 따른 Tg의 정보충분성 지표 ····· 28
그림 4. 표식유전자 정보유무에 따른 Tg의 정보충분성 순위 ····· 28
그림 5. 양적형질의 정보충분성 분포(ApoE) ····· 31
그림 6. 양적형질의 정보충분성 분포(AGT(2)) ····· 32

표 차 례

표 1. i 번째 가족의 잠재적 정보충분성 ····· 12
표 2. i 번째 가족의 유전 벡터 ····· 13
표 3. 표식유전자 ApoE의 자료특성 ····· 21
표 4. 표식유전자 AGT(2)의 자료특성 ····· 22
표 5. 4인가계의 연관성 분석 결과(ApoE) ····· 25
표 6. 양적형질의 범위(평균±표준편차) ····· 31
표 7. 정보충분성 지표를 이용한 연관성 분석 결과(ApoE) ····· 33
표 8. 정보충분성 지표를 이용한 연관성 분석 결과(AGT(2)) ····· 34

국문요약

양적형질유전자의 연관성 분석을 위한 가계자료의 정보충분성 연구

본 논문은 연관성 분석에 있어 검정력을 높이고 시간·비용적인 절감을 위한 방법의 하나로 정보충분성을 이용한 가족자료 선택에 대해 연구하였다.

양적형질이 조사된 경우 가정된 모형이 실제모형을 잘 반영한다는 가정을 바탕으로 각 가계의 잠재적 정보충분성을 양적 지표로 표현하였다. 이를 통해 양적형질이 조사된 가계자료를 이용하여 정보충분성 지표를 기준으로 순위를 주어 유전자 판독시 정보력 있는 가계를 선택하는 기준이 될 수 있다. 정보충분성 지표는 양적형질이 조사된 각 가계에서 나올 수 있는 모든 가능한 유전적 구성에서 실제 유전모형이 가정되었을 때 그 유전적 구성이 일어날 확률로 가중을 주어 계산한 χ^2 검정통계량의 합을 의미하게 된다.

본 방법을 이용하여 일 병원 심장혈관유전체연구센터의 실제자료에 적용하여 가계별 정보충분성 지표를 얻을 수 있었으며, 지표의 상위에 속하는 가계가 그렇지 않은 가계에 비해 연관성을 발견할 가능성이 높은 것으로 나타났다.

핵심되는 말 : 양적형질, 가족자료, 연관성 분석, 정보충분성, 분산성분방법

제 1 장 서 론

1.1 연구배경

20세기 후반 시작된 인간 게놈프로젝트(Human Genome Project : HGP)연구는 인간게놈의 염기서열을 결정하는 연구로부터 시작하여 인간의 DNA 서열 중 약 99%에 해당하는 부분을 판독하게 되었다.(Celera corp.) 이 결과를 통해 DNA 서열 자체와 상당수의 유전자의 위치파악은 이루어 졌으나, 보다 중요하게 생각되어져야 할 사항인 유전자별 기능 식별에 대한 연구는 아직 초보 단계이며, 대부분의 유전자는 그 기능이 알려져 있지 않아, 앞으로 그 위치와 기능과의 연관성(linkage)에 대한 연구가 더더욱 강조될 것으로 예측되어 진다.

연관성 분석은 이미 알려져 있는 표식유전자(marker)를 이용하여 규명하고자 하는 형질과 관련된 유전자의 위치를 추론해 나가는 것으로 이를 위해서는 자료에 포함되는 구성원 각각의 특성과 표식유전자의 판독(genotyping)이 필요하다. 하지만 구성원의 표식유전자를 얻는 과정은 장기간의 시간과 많은 비용이 소요되는 것이 일반적이다. 또한 장기간의 시간과 비용을 투자하여 표식유전자를 판독한다 해도 대부분의 형제자료 또는 가족자료의 경우 연관성 정보를 가지고 있는 자료의 비율이 매우 낮은 것으로 알려져 있다.(Purcell S., 2001) 때문에 유전자를 판독하기 전에 표식유전자와 형질의 연관성을 잘 보여 줄 수 있거나 혹은 연관성에 대한 정보를 많이 포함하고 있는 자료를 선별해서 표식유전자의 판독을 시행한다면 적은 수의 자료로 원하는 수준의 추론을 수행할 수 있게 되고, 이를 통해 시간 및 경제적인 측면에서의 효율성을 높일 수 있다고 할 수 있다. 형질과 표식유전자의 적절성을 평가하는 중요한 척도로서 자료의 정보충분성(informativeness)이 사용될 수 있다. 정보충분성은 조사된 자료의 단위(형제, 가계)가 함유하는 정보량을 의미하게 되며, 표식유전자 판독 전에 시행되게 된다.

기존의 연관성 분석을 위한 정보충분성의 측정에 관한 연구방법은 질적형질

자료에서 이루어져 왔고(박윤주 2001), 양적형질 자료에서는 형제자료(sib-pair)에 국한해서 특정 기준(threshold)에 해당하는 형제를 선택하는 방법에 따른 정보충분성에 대한 연구가 이루어져 왔다(Amos C.I., 1994; Carey G., Williamson 1991; Dolan C.V., Boomsma DI 1998; Zhang N., Zhang H. 1995). 형제자료의 선택방법은 표현형의 형질이 극값을 갖고 있는 개인을 정의하는 기준(threshold)을 사용하여 기준에 포함되는 자료에서 연관성(linkage)을 발견하는데 있다. 이러한 방법들은 자료의 특성에 따라 다른 기준이 적용되기 때문에(Purcell S., Cherny S.S., 2001) 어떠한 조건에서도 궁극적으로 사용되어질 수 있는 방법의 필요성 대두되었다. 이런 단점을 보완하여 S. Purcell은 어떤 조건에서도 정보충분성이 높은 형제자료를 선택하는 방법을 제시하였으며, 이는 분석에 사용된 가정이 실제모형(true model)을 정확히 반영한다고 가정할 때 기준을 사용한 형제선택 방법보다 좋은 방법이라고 알려졌다.(Purcell S., Cherny S.S, 2001) 하지만 아직까지 양적형질 가계자료의 정보충분성에 대해서는 논의 되지 않고 있으며, 가계자료의 특성에 상관없이 궁극적으로 사용될 수 있는 지표가 필요하다고 하겠다.

1.2 연구 목적 및 방법

어떤 자료가 효율성을 높을 수 있는 자료인지에 대한 평가는 양적형질 또는 질적형질 유전자를 발견하기 위한 검정력(power)을 높이는지에 대한 확인으로 판단할 수 있다. 일반적으로 검정력을 높이기 위해 표본 크기를 늘리는 방법을 사용할 수 있는데 이는 표식유전자의 판독비용을 증가시키는 제약을 가지고 있다. 이에 대한 대안으로 가장 잠재적 정보가 충분한 관측치를 선별하여 판독하는 방법이 대두되고 있으며 이의 중요한 척도로 정보충분성(Informativeness) 지표를 사용한다.

본 논문은 가계 구조가 조사된 자료를 이용하여 연관성 분석을 시행할 때 표식유전자 판독 전에 각 가계가 판독에 적절한지의 여부를 판단하기 위한 정보충

분성(informativeness) 지표를 측정하는 방법을 제시한다. 기존의 양적형질의 형제쌍에서만 가능했던 정보충분성의 의미를 형제 수에 상관없이 측정할 수 있는 기반을 마련한 연구(S. Purcell, S.S Cherny 2001)를 기반으로 QTL의 실제 유전(true genetic) 모델을 가정했을 때, 기존의 형제자료에서만 사용가능 했던 방법을 확장하여 일반가계자료에서 적용가능한 정보충분성의 척도를 제시한다.

현실적으로 가계의 표식유전자를 판독하는 것은 어려운 일이다. 때문에 가계도 구성 및 표식 유전자 유형이 전제되고 실제로 표식유전자 판독이 이루어진 경우의 실제자료를 이용하여 분석한다. 조사된 표식유전자 정보를 모른다는 가정 하에 분석하며, 후에 조사된 표식유전자를 사용하여 유전자 판독 과정을 대신한다. 실제 가족자료에서 표식유전자를 선택하고 표식유전자와 관련 있다고 알려진 양적형질을 이용한다. 실제모형을 가정하는 과정에서는 전체 자료를 기반으로 한 추정치를 사용하여 분석하였다.

논문에서는 우선 기존의 방법인 질적형질 가족자료의 정보충분성과 양적형질 형제자료의 정보충분성에 대해 소개하고, 양적형질 가족자료로 확장하여 논의한다. 실제자료를 이용한 분석을 통하여 정보충분성의 지표가 높게 나타난 그룹과 낮게 나타난 그룹 간에 연관성 분석결과의 차이가 있는지 알아본다. 동일한 자료에 대해 표식유전자가 조사된 경우와 모든 표식유전자가 결측된 경우의 정보충분성을 측정하여, 결측치가 정보충분성에 영향을 주는지에 대해 확인해 보도록 한다.

제 2 장 정보충분성

2.1 개요

지금까지 알려진 정보를 가진 형제자료 또는 가족자료를 선별하는 방법에 대해 논의해 본다. 이는 형질의 분류에 따라 또는 선택되어 지는 자료의 구성 형태에 따라 분류되는데 본 장에서는 질적형질 가족자료 선택방법과 양적형질 가족자료의 선택방법에 대해 논의해 보기로 한다. 질적형질 자료의 정보충분성을 구하는 방법은 유전적 모형에 기초한 모형기반 분석방법(model-based method)에 의해 구해지는 반면 양적형질 자료의 정보충분성을 구하기 위해서는 비전형적 유전형태를 가정하는 모형무관 분석방법(model-free method)이 사용된다. 양적형질의 유전자 연관성 분석방법 중 R. A. Fisher가 1918년에 제안한 분산성분방법을 사용하여 양적형질의 분산을 유전적 분산과 환경적 분산으로 나누어 각각의 값을 추정하고 검정하는 모형무관방법을 사용한다. 우선 기존에 많이 알려진 양적형질을 가진 형제자료의 정보충분성을 구하는 방법에 대해 소개한다.

2.2 일반가계자료의 질적형질 정보충분성

두 유전자의 연관된 정도는 유전적 거리(map distance)와 밀접한 관련이 있으며 유전적 거리를 이용하여 특성 유전자의 위치를 알아내는 것을 목적으로 한다. 세포의 분열과정 중 염색체의 교차(cross over)로 인한 염색체의 분절(segment)의 교환으로 새로운 조합의 대립유전자(allele)가지게 되는데 이를 유전자의 재조합(recombination)이라 한다. 유전자들의 재조합을 통해 부모에게서는 발견되지 않은 새로운 형태의 대립유전자 쌍을 만들어 유전시킨다. 이 때 유전자간 거리가 짧을

경우 교차되기 어렵기 때문에 교차횟수가 적고, 재조합률(recombination fraction)도 낮아지지만 유전자 거리가 멀수록 상대적으로 교차되기 쉽기 때문에 교차 횟수는 많아지고 재조합률도 높아지게 된다. 이런 특성을 이용하여 재조합률의 추정과 검정을 통한 연관성 분석을 시행할 수 있으며, 재조합률이 낮을수록 연관되어 있다고 말할 수 있다.

조사된 가계도의 정보를 F라하고 두 유전자 사이의 재조합률을 θ라고 정의 할 때, 재조합률을 추정하는 방법은 두 가지가 있다. 첫 번째 방법은 θ에 대한 우도함수를 $L(\theta|F)$과 같이 표현하여 가계도 정보가 주어졌을 때의 재조합률 θ의 확률에 의해 정의되는 우도함수(likelihood function)방법이고, 두 번째 방법으로는 θ의 사전 분포(prior distribution)를 가정하고 자료의 표본 분포를 통해 구해진 $L(F|\theta)$를 이용하여 θ의 사후 분포를 통한 θ의 재조합률을 추정하는 베이지안(Bayesian)방법이 있다.

이렇게 추정된 재조합률을 이용하여 연관성 검정을 시행하며, 일반적으로 두 유전자가 연관되어 있지 않다고 가정하는 귀무가설($\theta = 0.5$)과 두 유전자가 연관되어 있다는 대립가설($\theta < 0.5$)에 대해 검정하며, 연관되어 있다고 가정했을 때의 우도와 연관되어 있지 않다고 가정했을 때의 우도비의 log값인 *Lod Score*가 사용되며 그 식은 아래와 같다.

$$Z(\theta) = log_{10}\frac{L(\theta|F)}{L(\theta = 0.5|F)}$$

위의 식에서 계산된 *Lod Score*에서 θ를 최대화 시키는 값을 최대 *Lod Score* (maximum Lod Score)라고 하며,

$$Z_{\max} = Max.Lod = log_{10}\frac{L(\hat{\theta}|F)}{L(\theta = 0.5|F)}$$

으로 나타낼 수 있다. 이 때 $\hat{\theta}$는 최대우도 추정치(Maximum likelihood estimate : m.l.e)를 나타낸다.

이 값을 기반으로 각 재조합률 θ에서 유전자 재조합이 발생한 명수에 대한 *Lod Score*에 대해 유전자 재조합이 발생할 확률로 가중(weight)을 준 *Lod Score*의 가중평균으로 구한 *ELOD*를 이용하여 정보충분성을 검정할 수 있다.

$$ELOD = \sum_i P(i) Z_i(\theta), \qquad P(i) \text{ : 가중치}$$

*ELOD*는 하나의 가계도에서 계산된 값으로만 평가하는 것이 아니라 구조가 다른 가계도를 서로 비교하여 상대적인 정보충분성을 평가하므로 상대적 해석이 필요하다. *ELOD*와 함께 최대 *Lod Score*가 특정 상수보다 클 확률 $P(Z_{\max} \geqq c)$ 을 사용하기도 하는 데, 일반적으로 상수 c 는 3을 이용하지만 상황에 따라 더 낮은 값을 쓰기도 한다.

이 방법은 침투율(panetrance rate)이나 대립유전자의 빈도 등 유전적 모형에 대한 기본가정에 대한 의존도가 매우 큰 유전적 모형에 기초한 분석방법(genetic model-based method)이다. 때문에 자료의 특성에 따라 민감하게 반응하여 검정력에 문제가 있을 수 있으며, 실제로 유전모형에 대해 정확한 가정을 할 수 없는 경우가 대부분인 점을 가정하면 왜곡된 결론을 도출할 가능성이 크다.

2.3 형제자료의 양적형질 정보충분성

양적형질 형제자료의 선택방법은 표현형인 특성이 극값을 갖고 있는 기준(threshold)을 사용하여 개인을 정의하여 선택하는 방법으로 각 기준에 포함되는 형제자료에 있어 연관성(linkage)을 발견하는데 있다. 양적형질 형제자료에서 가장

큰 정보를 준다고 알려진 세 가지의 기준은 형제 모두 형질이 높거나(concordant high), 형제 모두 형질이 낮거나(concordant low) 또는 형제의 형질이 서로 상이한(discordant) 경우를 나타내며, 형제 쌍의 선택방법은 이 테두리 안에서 발전되어 왔다.

몇 가지 선택 기준에 대해 살펴보면, 질병에 걸린 형제 선택방법(affected sib pair design)이 있다. 이는 형제 모두 표현형 기준보다 특성이 더 높은 경우 선택하는 방법으로, 선택되어지는 형제자료는 대부분 형제의 상관관계에 영향을 받는다.

다른 방법으로는 프로벤드 선택방법(proband selection)이 있다. 발단자 선택방법이라고도 불리는 이 방법은 형제 쌍 중 적어도 한명이 표현형 기준보다 더 높아야 하며 나머지 한 형제는 동일하게 높거나 상이한 형제를 선택하는 방법이다(Carey G, Williamson J, 1991). 하지만 대부분의 경우 형제의 특성이 서로 극도로 상이한 경우(extremely discordant)가 형제의 특성이 모두 높거나 낮은 경우보다 더 정보력을 가지고 있다고 알려졌다. (Risch N, Zhang H, 1995; Risch N, Zhang H, 1996)

많이 사용되는 또 다른 선택방법은 형제의 특성이 극도로 상이하거나 형제의 특성이 모두 극도로 높거나 낮은 형제를 함께 선택하는 방법(extreme discordant and concordant)이다. 이 경우 좀 더 효율적인 표본 선택을 위하여 유전모델의 정보에 따라 극도로 높거나 낮은 기준의 정도를 조정할 수 있다. (Gu C, Todorov A,Rao DC, 1996)

위에서 말한 선택 방법들은 자료의 특성에 따라 적절한 방법과 그 기준이 바뀐다는 단점을 가지고 있다. 예를 들자면 주된 대립유전자(allele)의 빈도가 낮은 경우에는 형제의 특성이 극도로 높은 값을 갖는 표본을 선택하는 것이 가장 효율적인 선택방법이 되고, 주된 대립유전자의 빈도가 높은 경우에는 극도로 낮은 특성을 갖는 형제자료만을 선택하는 것이 효율적인 선택방법이 되는 것이다. (Risch N, Zhang H, 1996)

2001년 S.Purcell 은 어떤 조건에서도 성립되는 궁극적(optimal)인 정보충분성에 대해서 말했었다. 궁극적이라 함은 자료가 주어졌을 때 정보충분성 지표의 상위

5% 선택한다면 주어진 자료에서 어떤 조합의 자료를 추출하더라고 선택된 5% 보다 높은 정보를 가지고 있지 않다는 의미이다. 이 방법은 최대우도 분산성분법(maximum likelihood variance component)을 이용하였으며, 우도 함수는 유전적 변동에 대한 성분과 환경적 영향에 대한 성분으로 구성된 공분산 구조와 평균벡터에 대한 모수화에 영향을 받게 된다. 이를 통해 형제 자료 선택에 사용되어질 정보충분성 지표를 구할 수 있었으며, 자료의 구조나 유전모델의 가정에 상관없이 어떤 경우에도 다른 기준에 의한 선택 방법보다 검정력을 높이는 방법이라는 것을 보여줬다.

본 논문은 S.Purcell의 2001년 논문에서 논의 되었던 형제자료에서의 정보충분성 지표를 이용한 선택방법을 가족자료로 확장시켜 양적특성을 가진 가족자료의 정보충분성 지표를 구하기 위한 방법론을 논의한다. 아울러 방법의 평가를 위해 일 병원의 심혈관 자료를 이용하며 분석하며 그 결과를 해석, 토의 한다.

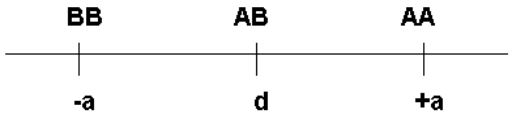
제 3 장 일반가계자료의 양적형질 정보충분성

3.1 개 요

이 장에서 제시하는 방법은 최대우도(maximum likelihood) 분산성분 방법(variance component method)을 기초로 한다. 분산성분에 기반을 둔 연관성 분석은 복잡한 양적형질에 영향을 주는 유전자를 알아내는 중요한 방법으로 형제자료뿐만 아니라 확장된 가계자료에서도 사용가능하며, 여러 유전자를 동시에 분석할 수 있는 장점이 있다. 전체분산은 유전적 요소와 가족간에 공유된 환경적 요소, 가족간 공유되지 않은 환경적요소의 세 가지 분산성분으로 이루어 졌다고 가정하며, 공분산 구조는 allele sharing IBD(identical by descent)의 영향을 받는다. 또한 가정한 실제모형이 자료의 특성을 잘 반영한다고 가정한다.

3.2 유전 모형

양적형질이 두개의 대립유전자(allele)를 가진 양적형질 유전자(quantitative trait loci : QTL)에 의해 영향 받는다고 하면, 완벽한 정보를 가진 표식유전자는 QTL에서 0 cM에 위치한다고 가정한다. 양적형질에 유전자는 A와 B 두개의 대립유전자를 가지고 두 대립유전자의 빈도를 p, $q = 1 - p$ 로 가정하면, 이 때 가능한 양적형질 유전자는 AA, AB, BB 세 가지 유전형으로 나타나고 그에 따른 유전적 효과는 a , d, $-a$으로 정의한다.



지배적 유전적 효과(dominance genetic value, d)와 가법적 유전적 효과(additive genetic value, a)의 비를 $z = d/a$, QTL에 의해 설명되어지는 표현형 분산의 비율을 x , 가족간 상관관계 r 을 가진 QTL로 정의된다.(Falcorner's 표기법) 가족의 형질 값을 기초로 정보충분성 지표의 기대값을 구하는 과정에서 각 가족의 상대적 순위에 영향을 미치지 않게 하기 위해 재조합률은 없다고 가정한다.

형질의 총 분산을 V_{Total} 이라고 하면 가법적 유전 효과와 지배적 유전 효과는 아래와 같이 표현된다.

$$a = \sqrt{\frac{xV_{total}}{(2pq(a+z(q-p))^2 + (2pq)^2}}$$

$$d = za$$

V_{Total}은 가법적 유전 효과로 설명되는 분산($V_{additive}$), 잔여 유전 효과와 공유하는 환경적 효과로 설명되어지는 분산(V_{Shared}), 잔여 유전효과 및 공유하지 않는 환경적 효과로 설명되어지는 분산($V_{Non-Shared}$)의 세 개의 분산성분으로 나누어지며 각각의 분산성분은 가법적 유전 효과와 지배적 유전효과 대립유전자의 빈도를 이용한 함수로 표현된다. 때문에 임의의 가계의 총변동이 결정이 되면 대립유전자의 빈도 등을 통해 가법적 유전 효과와 지배적 유전효과를 구할 수 있으며 이를 통해 임의의 가계의 분산성분들을 계산할 수 있다.

$$V_{Total} = V_{additive} + V_{Shared} + V_{Non-Shared}$$

$$V_{addtive} = 2pq(a + d(q-p))^2$$

$$V_{Shared} = rV_{Total} - \frac{V_{additive}}{2}$$

$$V_{Non-Shared} = (1-r)V_{Total} - \frac{V_{additive}}{2}$$

p와 q는 알려진 대립유전자의 빈도를 사용하거나 표본 또는 모집단의 빈도를 이용한다.

3.3 정보충분성 지표

본 연구는 실제 유전 모델을(true genetic model)을 가정할 수 있다고 할 때, 형질이 얻어진 가족자료를 이용하여 QTL 연관성 분석을 위한 가족자료의 잠재적 정보충분성의 지표를 구해 내는데 있다. 이를 위해 가족 내의 모든 구성원의 독립적인 공헌도의 합을 정의하여 이를 그 가족의 잠재적 정보충분성의 지표로 보기로 한다. 이 방법을 통하여 가족구성원의 수에 상관없이 비교 가능한 가족별 정보충분성 지표를 구해낼 수 있다.

공헌도 측정을 위해서 한 가족의 모든 가능한 유전적 구성(genotypic configuration: GC)들을 구해야 한다. 한 가족의 모든 가능한 유전적 구성에 대해 최대우도분산성분 방법을 이용하여 구한 각각의 검정통계량에 각 유전적 구성이 나올 확률로 가중을 주어 모든 검정통계량을 합함으로서 그 가족의 정보충분성의 지표가 계산되어 진다.

i 번째 가족의 표식유전자의 가능한 구성비를 n가지라고 하면 표 1.에서처럼 n가지 유전적 구성 각각의 확률을 계산할 수 있으며, 유전적 구성 각각의 검정통계량을 계산해 낼 수가 있다. i 번째 가족의 각 유전구성에 따른 검정통계량에 그 유전구성이 나올 확률로 가중을 준 값의 합(③)이 i 번째 가족의 잠재적 정보충분성의 지표로 사용된다. 이 방법을 사용하여 가족구성원의 수에 상관없이 구할 수 있으며, 이를 통해 서로 구성원의 수가 다른 가족들 사이에서도 정보충분성의 지표를 비교하는 것이 가능하다.

표 1. i 번째 가족의 잠재적 정보충분성

유전적 구성(GC)	$P(GC\|x)$①	χ^2 ②	①×②
1	P_1	t_1	P_1t_1
2	P_2	t_2	P_2t_2
3	P_3	t_3	P_3t_3
.....			
n	P_n	t_n	P_nt_n
			$\sum P_it_i$③

① 실제 유전모델이 가정되고, 양적형질이 주어졌을 때 계산되어지는 유전적 구성확률
② 검정통계량
③ i 번째 가족의 최종 정보충분성 지표

3.3.1 가계의 유전적 구성

연관성 분석에서 사용되는 가계의 유전적 구성(genotypic configuration;GC)은 부모의 유전자 구성에 기초하며, 유전모델에 의해 추론되어지는 유전형태로 구성될 수 있다. 이 때 부모의 유전자 구성 방법의 수는 QTL의 대립유전자의 수로 결정되게 되고, 전체 유전자 구성방법의 수는 부모의 유전자 구성방법과 대립유전자 수, 자식의 수에 영향을 받게 된다.

부모와 2명의 자식으로 구성된 핵가족을 예로 들어 보자. 만약 우리가 관심 있는 표식유전자가 1, 2, 3, 4의 대립유전자로 구성이 된다면, 모든 가능한 아버지와 어머니의 유전자 구성의 조합형태가 나타날 수 있고, 그 각 경우 자식이 받을 수 있는 유전자 형태가 정해지게 된다. 표를 보면 만약 부모가 각각 1/2와 3/4형태의 유전자를 가지고 있다면 그 부모의 유전적 구성에 대해 두 명의 자식이 가질 수 있는 유전자는 1/3, 1/4, 2/3, 2/4의 4가지가 되고, 두 명의 자식이 네 가지의 유전

자를 내려받아 구성하게 될 유전적 구성방법은 16가지가 된다. 자식의 수를 s 라고 하고 부모로부터 내려받을 수 있는 가능한 유전자의 형태가 2^2개라면 자식은 2^{2S}개의 유전자 구성을 가지게 된다. 하지만 실질적인 유전적 구성의 수는 부모가 가질 수 있는 유전적 구성의 수와 그 각각의 경우 자식이 가질 수 있는 유전적 구성의 수의 곱으로 나타난다.

이런 방법으로 구한 i 번째 가족의 각 유전적 구성은 IBD 행렬에 영향을 주며 각 구성별로 가족 구성원수 만큼의 차원을(위의 예에서는 4차원) 가진 IBD 행렬이 생성된다. 이 때 IBD는 유전적 일치도를 나타내며 한 쌍의 형제가 부모로부터 동일한 대립유전자를 내려받을 확률을 의미한다. f_i를 i개의 동일한 대립유전자를 갖고 있을 사전확률이라고 정의하면 IBD 비율의 추정치 π는 $\pi = f_2 + \frac{1}{2}f_1$으로 나타낼 수 있다. 표 2.에서와 같이 부모와 두 명의 자식으로 구성된 가계의 IBD는 아버지와 자식1, 아버지와 자식2, 어머니와 자식1, 어머니와 자식2, 자식1과 자식2 사이에서 계산되어 질 수 있으며, 부부는 서로 다른 부모를 가지고 있으므로 IBD한 대립유전자를 가질 수 없다. 첫 번째 유전벡터를 예로 들면 각각 부모와 자식간에는 하나의 공통된 대립유전자를 갖고 형제끼리는 두 개의 공통된 대립유전자를 갖는다.

표 2. i 번째 가족의 유전 벡터

	아버지		어머니		자식1		자식2		P(GCi)
	F	M	F	M	F	M	F	M	
유전벡터									
1	1	2	3	4	1	3	1	3	1/n
2	1	2	3	4	1	3	1	4	1/n
3	1	2	3	4	1	3	2	3	1/n
4	1	2	3	4	1	3	2	4	1/n
5	"	"	"	"	"	"	"	"	1/n
6	"	"	"	"	"	"	"	"	1/n
7	"	"	"	"	"	"	"	"	1/n
…	…	…	…	…	,,,	,,,	,,,	,,,	…
n									1/n

3.3.2 공분산 행렬

양적형질이 다변량 정규분포를 따르는 경우 분산성분방법을 이용한 연관성 분석에서 모든 QTL의 정보는 공분산 행렬에 표현되어 진다. 이 때 양적형질의 총 변동을 몇 가지의 분산성분들의 합으로 표현할 수 있다. 본 논문에서는 총 변동을 유전적 요소와 가족간에 공유된 환경적 요소, 가족간 공유되지 않은 환경적요소인 세 개의 분산성분으로 설명하고자 한다. 각 요소는 실제모형에 대한 가정에 의해 각 가족별로 계산되어 진다.

$$V_{Total} = V_{additive} + V_{Shared} + V_{Non-Shared}$$

i 번째 가계에서 g 번째 유전적 구성의 표식유전자가 연관되어 있다고 가정할 때의 공분산 행렬을 $\sum_{Lg}$, 연관되어 있지 않다고 가정할 때의 공분산 행렬을 $\sum_{Ng}$ 이라고 하면 각각의 공분산 행렬은 아래와 같이 표현된다.

$$\sum_{Lg} = \begin{bmatrix} V_{additive} + V_{Shared} + V_{Non-Shared} & \pi V_{additive} + V_{Shared} \\ \pi V_{additive} + V_{Shared} & V_{additive} + V_{Shared} + V_{Non-Shared} \end{bmatrix}$$

$$\sum_{Ng} = \begin{bmatrix} V_{additive} + V_{Shared} + V_{Non-Shared} & (1/2) V_{additive} + V_{Shared} \\ (1/2) V_{additive} + V_{Shared} & V_{additive} + V_{Shared} + V_{Non-Shared} \end{bmatrix}$$

가족의 구성원 수를 m 이라고 하면 가족의 공분산 행렬은 $m \times m$ 의 크기를 가지며 대각 원소는 분산성분들의 합인 총 변동을 나타낸다. 이 때 π는 대립유전자의 IBD비율을 나타내게 되는데 0, 1, 2개의 유전자를 내려 받을 확률은 각각 1/4, 1/2, 1/4 이고, 대각원소를 제외한 나머지 원소는 형제 또는 친척 간의 IBD 비율이 된다. 표식유전자가 연관되어 있다고 가정한 경우의 IBD는 가계자료의 유

전적 구성에 따라 구해지며, 연관되어 있지 않다고 가정한 경우의 IBD는 모든 경우에 공통적으로 1/2이 된다. 따라서 연관되어 있다는 대립가설 하에서의 공분산 행렬은 i 번째 가족의 g번째 유전적 구성방법에 의해 영향을 받으며, 연관되어 있지 않다는 귀무가설 하에서의 공분산 행렬은 모든 유전적 구성방법에 상관없이 i 번째 가족에선 동일하다.

이러한 분산성분 방법을 통한 연관성 분석은 가계자료를 하나의 단위로 생각하여 가족간의 비독립적 관계를 고려할 수 있는 특징이 있다. 때문에 다른 방법들에 비해 제1종 오류(type I error)를 안정적으로 유지할 수 있지만, 양적형질이 다변량 정규분포를 심하게 위배하는 경우 첨도(kurtosis)나 왜도(skewness)가 커져 오히려 제 1종 오류가 증가하는 경향이 있다. 이런 경우 정규성을 만족할 수 있도록 변환과정(transformation)이나 유사우도(quasi-likelihood)를 이용한 방법이 필요하다.

3.3.3 우도함수

앞에서의 공분산 행렬을 이용하여 가족자료의 우도함수를 구할 수 있는데 i 번째 가족의 형질이 정규분포를 따른다고 가정할 때 일반적 우도함수의 형태는

$$L = (2\pi)^{-m_i/2} |\sum_i|^{-1/2} e^{-1/2[(y_i-\mu_i)\sum_i^{-1}(y_i-\mu_i)]'}$$

으로 나타내며, μ_i 는 i 번째 가족의 평균, y_i 는 i 번째 가족구성원 각각의 형질(trait) 행렬, $\sum_i$는 i 번째 가족의 공분산 행렬이며, m_i는 i 번째 가족의 구성원 수이다. 공분산 행렬이나 평균벡터들은 직접 추정되거나, 또는 관심 있는 이론적 모수들의 함수에 의해 만들어 질 수 있다.

각 가족에서 표식유전자가 연관되어 있다고 가정할 때의 우도함수의 자연로그 값을 $ln(L_L)$이라하고, 연관되어 있지 않다고 가정할 때의 우도함수의 자연로그값을 $ln(L_N)$이라고 하면 i 번째 가족의 $ln(L_{Li})$, $ln(L_{Ni})$값은 아래와 같이 표현된다.

$$ln(L_{Li}) = -(\frac{m}{2})(2\pi) - \frac{1}{2}|\sum_{Li}| - \frac{1}{2}[(y_i - \mu_{Li})\sum_{Li}^{-1}(y_i - \mu_{Li})^T]$$

$$ln(L_{Ni}) = -(\frac{m}{2})(2\pi) - \frac{1}{2}|\sum_{Ni}| - \frac{1}{2}[(y_i - \mu_{Ni})\sum_{Ni}^{-1}(y_i - \mu_{Ni})^T]$$

여기에서 y_i는 i 번째 가족의 형질벡터이며, $\sum_{Li}$과 $\sum_{Ni}$는 각각의 공분산 행렬이 된다. μ_{Li}는 연관되어 있는 경우의 평균벡터이며, μ_{Ni}는 연관되어 있지 않은 경우의 평균벡터이며 다음과 같이 나타낸다.

$$\mu_{Li} = (X^T\sum_L^{-1}X)^{-1}X^T\sum_L^{-1}y_i$$

$$\mu_{Ni} = (X^T\sum_N^{-1}X)^{-1}X^T\sum_N^{-1}y_i$$

X는 1로 이루어진 m×1 열벡터이다.

위의 사항을 고려하여 유전적 연관성에 대한 검정을 하기위해 우도비 검정방법(likelihood ratio test)을 적용하는데 이론적 모수들은 관심 있는 모수가 포함된 모델을 적합시켜 얻어진 양적형질의 우도함수의 자연로그 값($ln(L_L)$)과 이 모수들이 포함되지 않았을 때 적합시킨 우도함수의 자연로그 값($ln(L_N)$) 즉, 특정 모수의 값이 영이라는 귀무가설 하에서의 우도함수의 자연로그 값을 얻어 이론적

모수들의 통계적 유의성을 검정 할 수 있다. 본 연구에서는 유전적 효과의 분산을 0으로 설정한 모형이 $ln(L_N)$이 된다. 표본의 수가 커지면 $2[ln(L_L) - ln(L_N)]$ 는 검정되는 모수의 개수를 자유도로 갖고 χ^2 분포를 근사적으로 따르는 검정통계량이 된다. 위와 같은 과정은 가족자료에서 얻어진 양적형질을 모델화하는 일반적 접근이라 할 수 있다. (Fulker D. W., 1999)

3.3.4 가중치

i 번째 가족의 n가지의 유전적 구성방법(GC)에 대해 구해진 n 가지 우도함수의 자연로그 값의 차는 각 유전적 구성방법이 나타날 확률로 가중이 주어지게 된다. $P(GC_g|\,x)$를 형질이 주어졌을 때 총 n개의 유전적 구성방법 중 g번째 유전적 구성방법이 나타날 확률이라고 하면 이는 형질이 주어졌을 때의 g번째 유전적 구성방법의 사후확률을 이용하여 베이지안 방법을 이용하여 다음과 같이 구할 수 있다.

$$P(GC_g|\,x) = \frac{P(GC_g)f(trait|\,GC_g)}{\sum_{g=1}^{n} P(GC_g)f(trait|\,GC_g)}$$

이 때 양적형질은 다변량 정규분포를 따른다는 가정 하에서 우도비 검정이 진행되므로 $f(trait|\,GC_g)$는 $N(\mu_g, \sum_g)$를 따르게 되며, i 번째 가족의 g번째 유전적 구성이 주어진 경우 $f(trait|\,GC_g)$는 다음과 같이 구할 수 있다.

$$f(trait|\ GC_g) = \frac{1}{(2\pi)^{m/2}} \frac{1}{|\sum_g|^{1/2}} exp[-\frac{1}{2}(y_g - \mu_g^*)\sum_g{}^{-1}(y_g - \mu_g^*)^T]$$

이 때의 μ^* 은 $(X^T\sum_{Ng}^{-1}X)^{-1}X^T\sum_{Ng}^{-1}y_i$ 으로 구해지며 X는 1로 구성된 m×1 열벡터가 된다.

모든 가능한 유전적 구성의 수를 n이라고 하면 $P(GC_g) = 1/n$으로 계산되며 $P(GC_g)$와 $f(trait|\ GC_g)$의 값을 이용하여 각 유전적 구성에 대한 가중치를 구할 수 있다.

3.3.5 최종 정보충분성 지표(informativeness index)

i 번째 가족의 유전적 구성의 수를 n가지라고 할 때, i 번째 가족의 정보충분성의 지표는 n가지의 유전적 구성에서 각각 계산되어진 우도비 검정방법의 검정통계량과 그 때의 가중치를 곱한 값을 합한 값으로 계산되어 진다. 위의 계산을 적용하여 구한 i 번째 가족의 정보충분성은 아래와 같이 나타낼 수 있다.

$$E_i(\chi^2|\ x, model) = \sum_{g=1}^{n} 2\,[lnL_{Lg} - lnL_{Ng}]P(GC_g|\ x)$$

제 4 장 실제자료를 이용한 정보충분성 측정

4.1 자료와 방법

4.1.1 표식유전자 및 양적형질의 선택

심혈관계질환 유전체연구센터에서 수집된 가계도 자료를 이용하여 정보충분성 지표를 계산한다. 표식유전자는 ApoE(Apolipoprotein E)와 AGT(2)를 사용했다. AGT(2) 유전자는 angiotensinogen이란 혈압과 심장 기능을 조절하는 물질을 생산하기 때문에 AGT(2)가 과활성화되면 고혈압 및 심장병을 유발하게 된다고 알려져 있으며, ApoE는 혈중지질대사에 관여한다고 알려져 있다.

두 개의 표식유전자 중 ApoE는 19번째 염색체에 위치하는 유전자로 E2, E3, E4의 세 가지 형태의 대립유전자를 가지고 있으며, 유전형은 E2/E2, E3/E3, E4/E4, E2/E3, E3/E4, E2/E4 의 6가지 형태를 가진다. 이 중 E4 대립유전자를 가진 사람이 그렇지 않은 사람에 비해 동맥경화 등의 심혈관계 질환에 걸릴 위험이 높은 것으로 알려져 있다. 본 분석에서는 두 개의 대립유전자를 갖는다고 가정하였기 때문에 E4를 제외한 나머지 대립유전자를 F라고 가정하면 E4/F, E4/E4, F/F의 세 가지 형태의 유전형을 가진 표식유전자로 간주하여 분석한다.

분석의 대상이 된 양적 형질은 심혈관계질환들과 관계가 있다고 알려진 BMI와 혈액 내의 지질 성분 네 가지를 이용하였다. 양적형질로 사용된 혈액내의 지질 성분은 총 콜레스테롤(total Chlesterol; Tchol), 고밀도 지단백(high density lipoprotein; HDL), 저밀도 지단백(low density lipoprotein; LDL), 중성지방(triglyceride; Tg) 이다.

4.1.2 자료의 형태

표식유전자와 양적형질의 각 조합에 포함된 입력 자료는 가족관계 정보가 포함되어 있으며, 양적형질이 모두 조사된 자료이다. 자료의 구성은 가족 아이디와 개인 아이디, 부모정보, 양적형질, 표식유전자를 포함하고 있다. 양적형질이 모두 조사된 가계에 한해 표식유전자가 모두 조사된 가계와 일부 또는 모두 결측인 가계가 포함된다. 모든 분석은 표식유전자와 양적형질들 간의 각 조합에서 가족구성원의 양적형질이 모두 조사된 가계만을 선별하여 분석한다. 때문에 각 분석에 사용된 자료는 가계수나 가계구성에서 모두 일치하지는 않을 수 있다. 예를 들어 BMI는 조사되었으나 LDL수치는 조사되지 않은 경우 그 가계는 BMI분석에는 포함되나 LDL 분석에서는 제외되게 된다.

총 524가계, 2132명에 대해 조사하였으며 위의 조건에 해당하는 가계를 선별하여 분석하였다. 각 표식유전자와 형질의 조합별 자료의 특징과 각 조합의 가족관계 분포는 표 3.과 표 4.에서 확인할 수 있다. 대략적으로 각 표식유전자 내에서는 형질별 성비나 세대수의 비율(2세대와 3세대)는 유사한 것으로 보여진다.

표 3. 표식유전자 ApoE의 자료특성

		BMI	HDL	LDL	Tchol	Tg
가계수		308	254	240	306	306
	2세대(%)	293(0.95)	244(0.96)	230(0.96)	290(0.95)	290(0.95)
	3세대(%)	15(0.05)	10(0.04)	10(0.04)	16(0.05)	16(0.05)
개인수		1218	968	917	1210	1211
	남(%)	623(0.51)	496(0.51)	466(0.51)	616(0.51)	617(0.51)
	여(%)	595(0.49)	472(0.49)	451(0.49)	594(0.49)	594(0.49)
가족관계 분포	부모자식	1164	898	852	1154	1156
	형제쌍	350	245	230	344	345
	여자-여자형제	107	73	73	108	108
	남자-남자형제	77	52	48	75	75
	남매	166	120	109	161	162
	조부모-손녀	60	32	34	64	64
	삼촌	55	22	21	53	53
	사촌	9	2	2	9	9

표 4. 표식유전자 AGT(2)의 자료특성

AGT(2)		BMI	HDL	LDL	Tchol	Tg
	가계수	183	133	122	179	179
	2세대(%)	168(0.92)	122(0.92)	113(0.93)	166(0.93)	164(0.92)
	3세대(%)	15(0.08)	11(0.08)	9(0.07)	13(0.07)	15(0.08)
	개인수	760	537	488	735	745
	남(%)	394(0.52)	278(0.52)	239(0.49)	379(0.52)	385(0.52)
	여(%)	366(0.48)	259(0.48)	235(0.51)	356(0.48)	360(0.48)
가족관계 분 포	부모자식	748	518	460	720	734
	형제쌍	236	149	127	225	232
	여자-여자형제	70	44	44	67	70
	남자-남자형제	53	30	21	50	51
	남매	113	75	62	108	111
	조부모-손녀	60	36	30	52	60
	삼촌	55	24	19	48	53
	사촌	9	2	2	9	9

4.2 분석방법

표식유전자의 결측치가 포함된 가계자료는 동일한 가계에 대해 모든 가능한 유전자 구성이 조합된 가계자료로 만들어 지며, 가계자료의 유전적 구성이 타당한지에 대해 검토한다. 후에, 가족자료의 IBD 행렬을 이용하여 각 가족별 유전적 구성에 대한 공분산행렬을 계산해 낸다. 이를 이용하여 각 가족의 모든 유전적 구성에서 연관이 되었다고 가정했을 때의 우도와 그렇지 않을 때의 우도비의 자연로그 값을 구한다. 이 검정통계량에 양적형질에 가중을 준 가중치를 곱한 뒤 합산하는 방법으로 한 가족의 정보충분성 지표를 얻게 된다. 이 때의 분산성분들은 전체 표본의 유전자 빈도를 p, q 로 하여 가족별로 분산성분을 계산하여 사용한다. ApoE의 경우 $p = 0.092608$, $q = 0.907392$인 표본빈도를 이용하였으며 지배적 유전성분은 없다($d = 0$)고 가정하였다. 가족간 상관정도는 $r = 0.25$일 때 각 분산성분 계산식을 이용하여 각 가족별로 분산성분을 계산하였다. AGT(2)의 경우 $p = 0.183833$, $q = 0.816167$, $d = 0$, $r = 0.25$으로 각 가족별 분산성분을 계산하였다.

R(http://www.r-project.org)으로 전체적인 프로그래밍을 하였으며, IBD행렬을 구축하기위해 R의 KINSHIP 라이브러리를 이용하였다. 각 가족별 유전적 구성별의 IBD 행렬은 SOLAR(by Blangero et. al., http://sfbr.org/sfbr/public/software/solar/inde x.html)를 이용하여 계산하였으며, SAGE 5.0 (http://darwin.cwru.edu/index.php)을 통해 전체 가족의 표식유전자 검정과 가족정보에 대한 분석을 시행하였다.

본 방법은 표식유전자가 조사되기 전인 양적형질만 조사된 상태에서 정보충분성이 높은 가족자료를 선별해 내는데 본래 목적이 있다. 표식유전자 ApoE 자료에서 부모와 두 명의 자식으로 구성되고 표식유전자가 모두 조사된 4명의 2세대 가족자료만을 선택하여 표식유전자를 모두 결측치로 처리하였을 때의 정보충분성 지표를 구한다. 그 지표를 기준으로 상위 20%와 하위 20%의 가족자료 각각에서 조사된 표식유전자 정보를 이용하여 연관성 분석을 시행한다. BMI, HDL, LDL,

Tchol의 양적형질을 이용하였다. 형질들 중 4인가계가 아닌 다른 가계형태에서도 완전한 표식유전자 정보가 많은 Tg를 이용하여 3인 가계에서부터 8인 가계의 자료를 이용하여 위의 분석을 시행하였다.

실제로 얻어질 수 있는 자료형태를 반영하기 위하여 실제자료인 심장혈관질환 병원의 자료를 이용한다. 이 자료에는 표식유전자의 완전한 정보를 가지고 있는 가계와 부분적인 정보를 가진 가계가 포함되어 있다. 두 개의 표식유전자와 다섯 개의 양적형질로 이루어진 10개의 조합에서 정보충분성 지표를 기준으로 상위 10% 하위 10%의 속하는 20개의 가계자료를 추출한다. 20개의 자료를 가지고 각각의 경우에 연관성 분석을 시행하여 정보충분성 지표가 상위에 속하는 가계와 하위에 속하는 가계에서 연관성의 차이가 있는지 확인하고 유전율의 유의성에 대해서 검토한다.

4.3 분석결과

4.3.1 정보충분성을 이용한 가계 선별

부모와 2명의 자식으로 구성되고 표식유전자에 대해 완전한 정보를 가지고 있는 가계자료를 이용하여 정보충분성 지표를 계산하였다. 총 자료 중 조건을 만족시키는 가계를 선별하였으며, 4인 가족에 제한을 두어 총 분석 가능한 가계 중 120~166가계의 분석을 시행했다. 각 양적형질에 대한 요약은 [표 5]에서 확인할 수 있다. 모든 양적형질에서 정보충분성 지표의 상위 20%, 하위 20%에 속하는 자료의 연관성 분석을 시행하였다. 분석결과([표 5])를 보면 정보충분성 지표의 상위 그룹이 하위 그룹에 비해 높은 LOD 값을 가졌다. 유전율에서도 상위 그룹이 더 높았으며, 모든 상위그룹에서 유의하게 나왔다. 선별된 4인 가계의 대립유전자 빈도는 $p = 0.499$ 으로 표본의 빈도와 큰 차이를 보였음에도 불구하고 표식자 정보 유무에 관계없이 상위가 하위보다 높은 LOD 경향을 보였다. 예로 HDL의 실제 표식유전자 정보를 이용했을 때와 본 방법에 의해 구한 지표를 비교하여 그린 그림 1.과 그림 2.를 보면 정보충분성 지표는 물론 정보충분성 지표의 순위에서도 상당히 일치됨을 확인 할 수 있으며, 다른 양적 형질에서도 비슷한 양상을 보였

표 5. 4인 가계의 연관성분석 결과(ApoE)

		가계수	유전율	LOD
BMI	상위 20%	33	0.667*	0.2632
(23.31±3.33)a	하위 20%	33	0.210*	0.0000
HDL	상위 20%	24	1.000*	0.6948
(45.85±11.80)	하위 20%	24	0.300*	0.0000
LDL	상위 20%	20	1.000*	0.9110
(118.80±39.27)	하위 20%	20	0.224	0.2343
Tchol	상위 20%	32	1.000*	0.7866
(194.39±44.5)	하위 20%	32	0.183	0.1040

a (평균± 표준편차)
* p-value < 0.05

다.

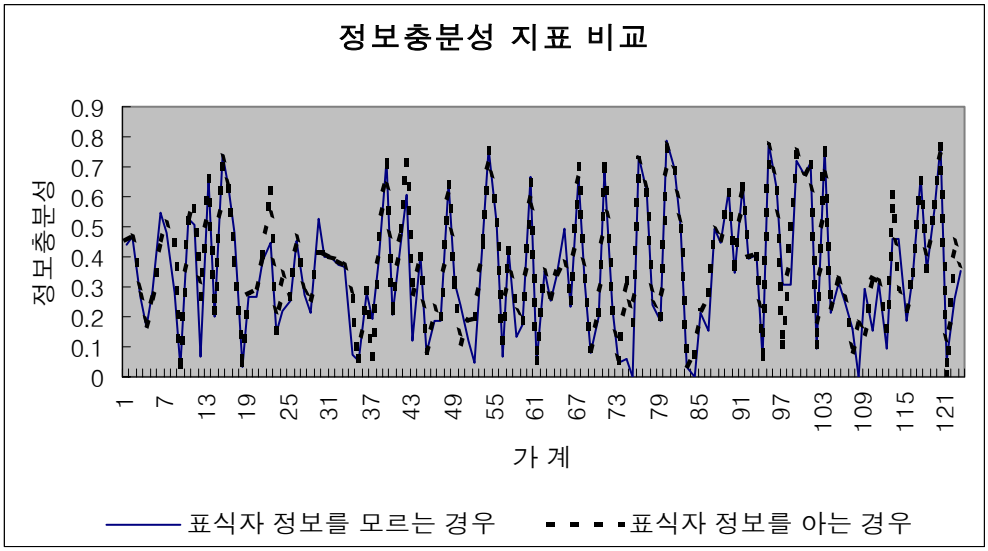


그림 1. 4인 가계에서 표식유전자 정보 유무에 따른 HDL의 정보충분성 지표

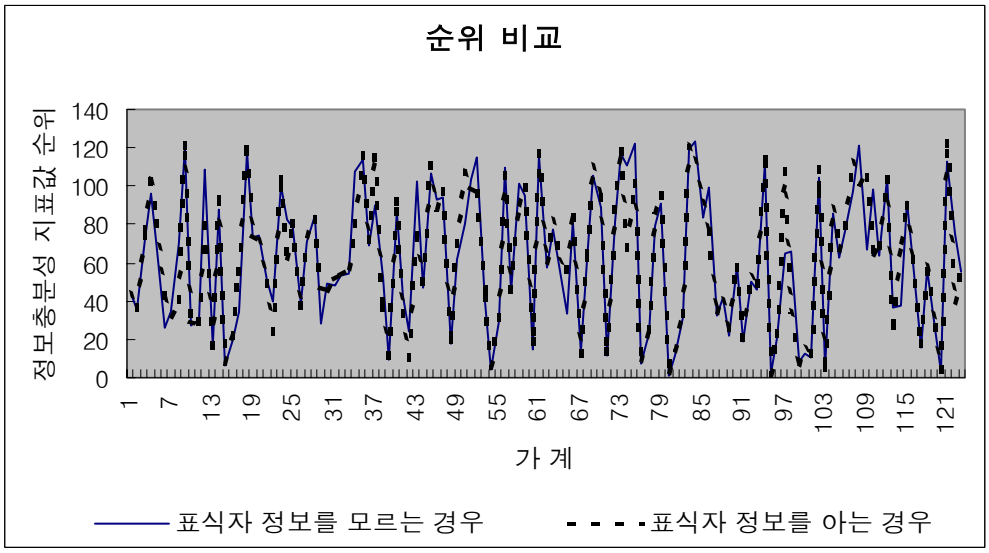


그림 2. 4인 가계에서 표식유전자 정보 유무에 따른 HDL의 정보충분성 순위

위의 방법이 4인 가계에 한정된 결과임을 고려하여 다양한 가족구성원 수를 가진 가계가 함께 있는 자료인 경우 본 방법과 실제 표식유전자 정보를 이용했을 때의 정보충분성의 차이를 알아보았다. 조사된 형질 중 비교적 표식유전자의 완전한 정보를 가진 가계가 많은 양적형질 Tg를 이용하여, 모든 표식유전자의 정보가 없다고 가정하고 구한 정보충분성 지표와 실제 완전한 정보를 가진 정보충분성 지표를 비교해 보았다. 모든 표식유전자가 조사된 147가계가 사용되었으며, 분석에 사용된 가계의 구성원 수는 3명에서 8명 사이의 분포를 가지고 있다. 실제 표식유전자로 구한 정보충분성 지표는 0.64~4.69의 분포를 가지고 있으며, 표식유전자의 정보가 없다고 가정하고 구한 정보충분성 지표는 1.058~6.445로 전체적으로 표식유전자 정보가 없다고 가정했을 경우 더 높은 경향을 보였다. 이는 그림 3.에서도 전체적으로 실제 정보충분성이 낮은 것으로 보인다. 그림 4.는 두 가지 경우 각각의 순위를 비교하였다. 그림 4.에서 상당히 일치한 결과를 확인할 수 있으며, 실제로 정보충분성 지표의 상위 10%에 속하는 15가계 중 13가계가 상위 10%에 속했으며, 상위 20%의 가계에 속하는 30가계 중 27가계가 포함되어 90%가 일치하였다. 상·하위 20%에 속하는 각 그룹에 대한 연관성 분석결과 상위 20%의 경우 LOD값이 1.0236이었으며 하위 LOD값은 0.0003이었으며 상위그룹은 유전율이 0.8634로 유의했으며, 하위 그룹은 0.2308로 유의하지 않았다. 결과에서 두 가지 방법의 정보충분성이 비슷한 값을 가지지는 않았지만, 표식유전자가 조사되기 전 정보충분성 지표로 가계자료를 선택한다고 할 때 실제 표식유전자의 판독을 통해 연관성을 발견할 가능성은 높다고 보여 진다.

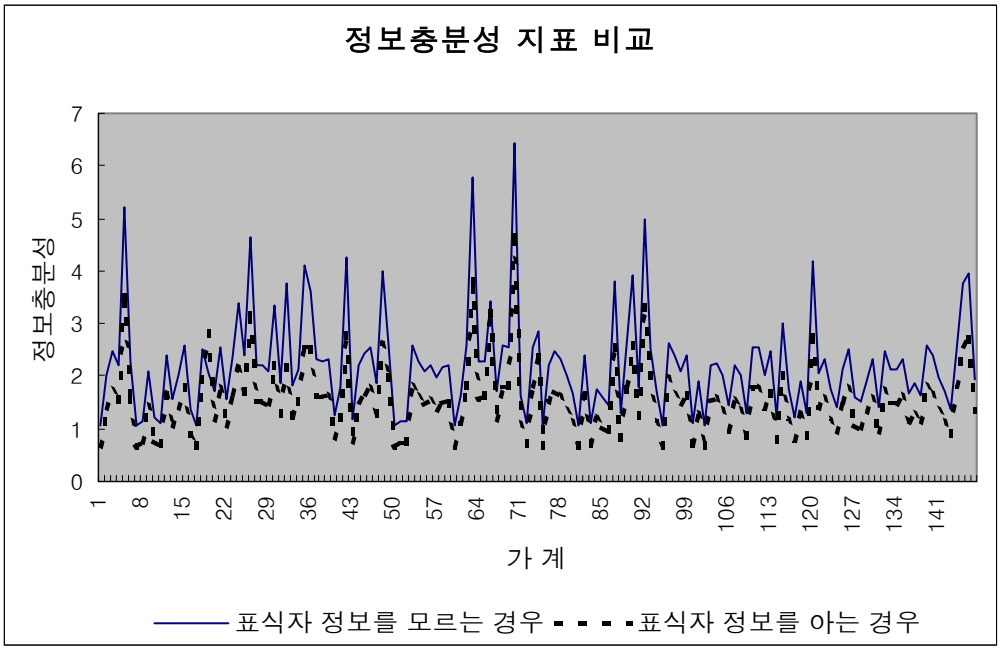


그림 3. 표식유전자 정보유무에 따른 Tg의 정보충분성 지표

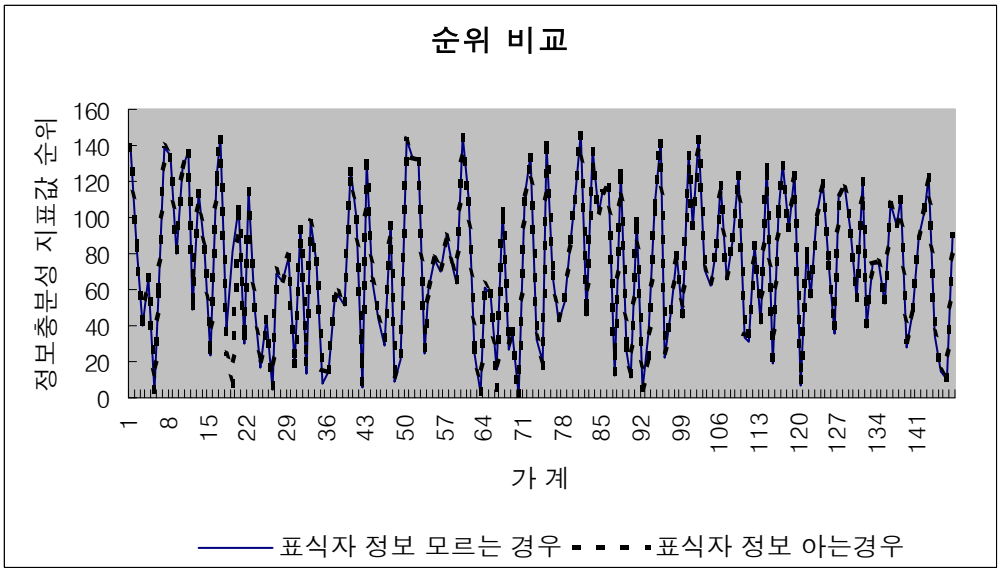


그림 4. 표식유전자 정보 유무에 따른 양적형질 Tg의 정보충분성 순위

4.3.2 전체자료의 정보충분성 측정

위의 결과로 표식유전자의 정보를 이용하여 구한 정보충분성의 지표와 표식유전자의 정보를 이용하지 않은 정보충분성의 지표 값 자체는 일관된 차이를 보이지만, 지표의 순위는 거의 유사한 경향을 보이는 것을 확인했다. 이를 바탕으로 각 형질별 전체 자료를 통해 정보충분성 지표를 구해보았다. 분석에 사용된 자료는 앞의 표 3., 표 4.에서 제시된 모든 가능한 자료를 활용하였다. 이 자료에는 완전한 정보를 가진 가계와 부분적인 표식유전자의 결측치를 가진 가계, 또 완전한 표식유전자 정보의 결측인 가계가 함께 포함된다.

각 표식유전자의 양적형질 별 정보충분성 지표 값의 분포는 그림 4., 그림 5.에서 확인할 수 있다. 대게 비슷한 양상을 보이지만 ApoE가 AGT(2)보다 높은 지표를 가지는 것을 확인할 수 있다. 각 조합의 정보충분성 지표를 기준으로 상위 10% 하위 10%에 속하는 가계를 추출해 각각 상위 10%와 하위10%로 정의한다. 2개의 표식유전자와 5개의 양적형질에서 각각 상·하위로 정의된 자료를 이용하여 각 경우의 자료의 특성, LOD값, 유전율(heritability)과 그 유의성에 대해 확인해 보았다.

표식유전자 ApoE에서 각 양적형질들에 대해 살펴보면, 각 양적형질들의 평균과 표준편차에서는 상·하위 그룹 간 특정한 경향은 보이지 않았다. 정보충분성 지표가 상위 10%에 속하는 그룹은 하위 10%에 속하는 그룹에 비해 전체적으로 평균가족수가 더 많은 것으로 나타났으며 평균세대수도 더 많은 것으로 보여 진다. 특이할 것은 대부분의 하위 그룹에 속하는 가계는 가족수가 3~4인 2세대인 경우가 대다수였다. SOLAR를 이용하여 분석한 유전율과 LOD 값에서는 상위 10%와 하위 10%간 차이를 확실하게 볼 수 있다. 상위 10%에 속하는 그룹은 하위 10%에 속하는 그룹에 비해 대게 유전율이 현저하게 높음을 확인할 수 있었으며, 통계적으로도 유의한 결과를 보였고, 하위 10%에 속하는 그룹은 유전율이 영이거나 거의 영에 가까운 수치를 나타냈으며 유의하지 않은 결과를 나타냈다. 연관성 분석에서 연관되어 있다고 판단할 수 있는 LOD 값 3인 기준에서 보면 3을

넘는 LOD 값은 Tchol의 상위 10%에서 나타난 5.7694의 경우 밖에 없었지만 다른 양적형질에서도 상위 10%에 속하는 그룹이 하위 10%에 속하는 그룹보다 LOD 값이 높음을 확인 할 수 있다. Tchol의 경우 그림 5.에서 Tchol의 정보충분성 지표가 가장 높았다.

표식유전자 AGT(2)에서 각 양적형질들에 대해 살펴보면, 정보충분성 지표가 상위 10%에 속하는 그룹은 하위 10%에 속하는 그룹에 비해 전체적으로 평균가족수가 더 많은 것으로 나타났으며 평균세대수도 더 많은 것으로 보여지는데 ApoE인 경우와 유사한 형태를 보인다고 할 수 있다. 하지만 위의 정보충분성에서도 ApoE보다 대게 낮은 값을 가졌기 때문에 LOD값도 대체적으로 낮고 상하위의 차이도 적었다. SOLAR를 이용하여 분석한 유전율(heritability)과 그 유의성에 대해서는 ApoE인 경우와는 조금 다른 양상을 보였는데 유전률이 거의 0에 가까웠던 ApoE에 비해서 AGT(2)의 경우는 하위 10%에서도 유전율이 어느 정도의 수치를 가진 값을 가지고 있는 것을 확인 할 수 있었다. 하지만 ApoE와 마찬가지로 상위 10%에 속하는 그룹은 하위 10%에 속하는 그룹에 비해 대게 유전율이 높으며 이는 통계적으로도 유의한 결과를 보였다. LOD 값은 모든 양적형질에서 상위 10%에 속하는 그룹이 하위 10%에 속하는 그룹보다 높게 나왔지만 3을 넘는 값은 나타나지 않았다.

위의 결과를 정리해 보면, 각 조합별로 약간의 차이는 있지만 본 논문에서 제시한 정보충분성의 지표로 분류할 때 지표의 상위에 속하는 그룹이 그렇지 않은 경우보다 연관성 분석에서 더 유의한 결과를 도출할 수 있을 것이라 기대할 수 있다. 또한 세대가 2세대 이상이고, 가족수가 많은 경우가 2세대의 핵가족 보다는 정보충분성이 높아 질 수 있다는 점을 확인 할 수 있었다.

표 6. 양적형질의 범위(평균(±표준편차))

	ApoE		AGT(2)	
	상위 10%	하위10%	상위 10%	하위 10%
BMI	22.69(±3.56)	23.19(±3.63)	23.26(±3.42)	21.93(3.45)
HDL	47.74(±12.87)	49.12(±11.91)	45.96(±11.70)	46.0(±10.85)
LDL	115.98(±32.42)	117.29(±29.76)	117.22(±32.74)	115.62(±32.68)
Tchol	182.67(±34.74)	193.96(±40.16)	188.48(±36.67)	205.63(±49.13)
Tg	125.81(±100)	140.46(±94.28)	158.0(±173.27)	149.08(±158.21)

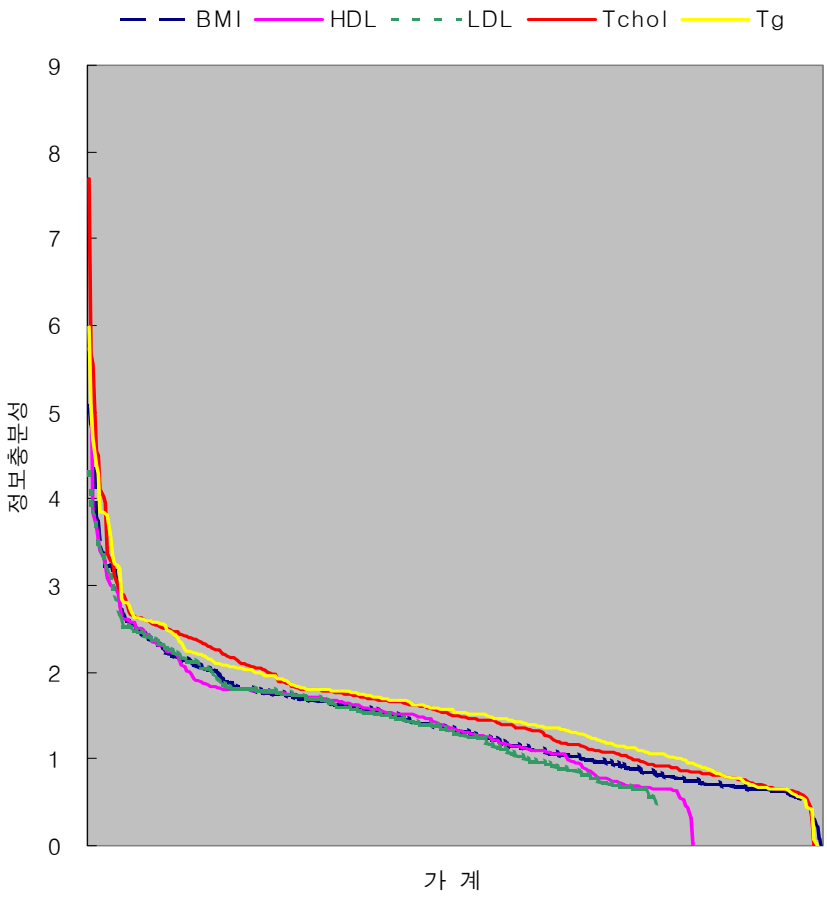


그림 5. 양적형질의 정보충분성 분포(ApoE)

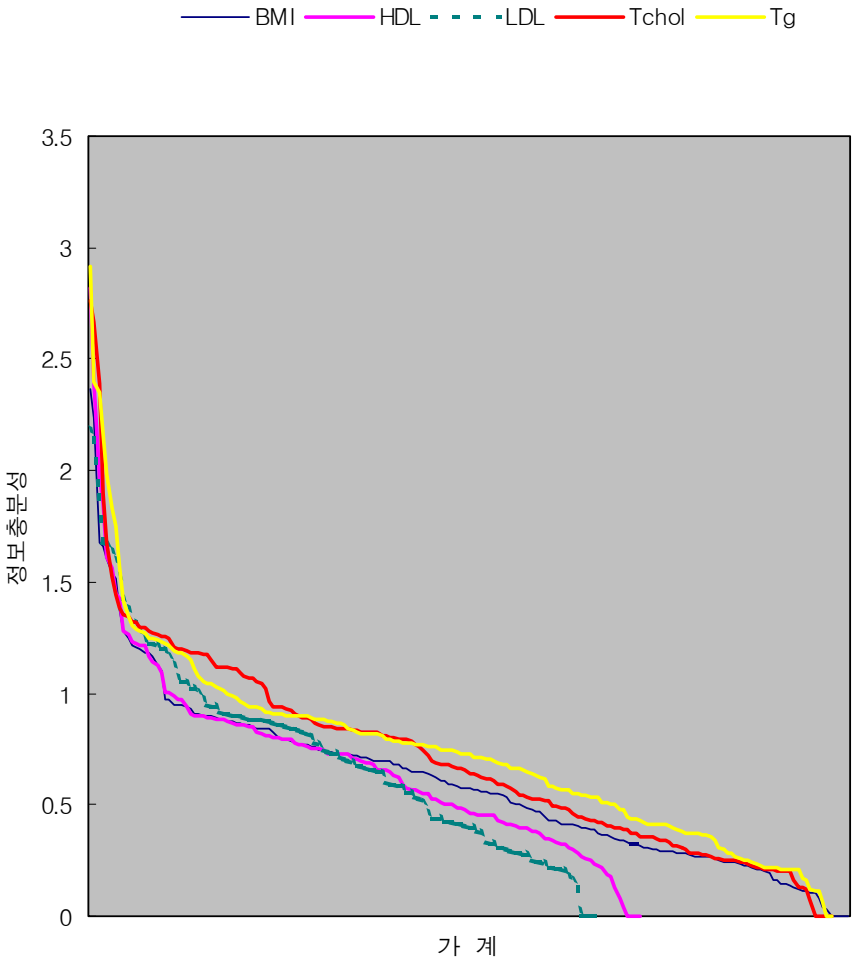


그림 6. 양적형질의 정보충분성 분포(AGT(2))

표 7. 정보충분성 지표를 이용한 연관성 분석 결과(ApoE)

양적형질		가계수	평균가족수	평균세대수	유전률	LOD
BMI	상위 10%	30	6.20(2.30)	2.43(0.50)	0.4750988*	0.4106
	하위 10%	30	3.90(0.44)	2.00(0.00)	0.0763999	0.0000
HDL	상위 10%	25	5.48(2.26)	2.28(0.21)	0.8588119*	0.0693
	하위 10%	25	3.44(0.42)	2.00(0.00)	0.0000000	0.0000
LDL	상위 10%	24	5.21(1.59)	2.33(0.70)	0.9389982*	1.2380
	하위 10%	24	3.67(0.92)	2.04(0.24)	0.0000000	0.0000
Tchol	상위 10%	30	6.27(2.27)	2.43(0.50)	0.7087675*	0.0187
	하위 10%	30	3.73(0.96)	2.03(0.03)	0.0000000	0.0000
Tg	상위 10%	30	6.23(2.18)	2.40(0.50)	0.5376780*	5.7694
	하위 10%	30	3.57(0.53)	2.00(0.00)	0.0000000	0.0000

표 8. 정보충분성 지표를 이용한 연관성 분석 결과(AGT(2))

양적형질		가족수	평균가족수	평균세대수	유전률	LOD
BMI	상위 10%	18	6.17(2.66)	2.44(0.51)	0.6897246*	0.0228
	하위 10%	18	4.50(0.98)	2.06(0.24)	0.0000000	0.0000
HDL	상위 10%	13	6.23(2.17)	2.53(0.52)	0.4380287*	0.0637
	하위 10%	13	4.15(0.99)	2.00(0.00)	0.2368627	0.0000
LDL	상위 10%	12	5.67(1.56)	2.33(0.49)	0.7358373*	0.0825
	하위 10%	12	4.00(0.95)	2.08(0.29)	0.1711325	0.0000
Tchol	상위 10%	17	6.65(2.50)	2.47(0.51)	0.6252775*	0.1973
	하위 10%	17	3.82(0.81)	2.06(0.24)	0.2590384	0.0000
Tg	상위 10%	18	6.56(2.57)	2.56(0.51)	0.5369100	0.0000
	하위 10%	18	3.94(1.66)	2.06(0.24)	0.0000000	0.0160

제 5 장 결론 및 고찰

지금까지 양적형질의 가족자료에서 정보충분성 지표를 구하기 위한 방법에 대해 알아보았다. IBD의 비율을 이용한 분산성분방법을 이용하여 가계의 정보충분성 지표를 구했으며 구성원 수에 상관없이 비교 가능한 방법으로 그 활용도가 높다고 할 수 있다.

4인 가계인 경우 모든 가계의 표식유전자를 모른다고 가정하고 정보충분성을 구했을 때와 표식유전자 정보를 이용하여 정보충분성 지표를 구한 각 정보충분성 지표에 따라 상·하위 20%의 가계를 선별하여 연관성 분석을 시행한 결과 상위 그룹에서 좀 더 유의한 결과를 보였으며, 유전율에서도 하위 그룹에 비해 높은 값을 가지는 것으로 분석되었다.

좀 더 다양한 가계형태를 반영하여 3인~8인의 가계가 모두 포함된 경우 정보충분성 지표를 구해 비교했을 때에도 4인 가계를 이용한 분석과 유사한 결과를 얻을 수 있었다. 전체적으로 표식유전자를 이용했을 때가 이용하지 않았을 때보다 정보충분성 지표가 낮게 나타났다. 하지만 각 경우의 정보충분성 지표를 기준으로 순위를 부여하여 비교한 결과 순위에는 큰 차이를 보이지 않았다. 즉 만약 연구의 목적이 정보력을 가진 상위 그룹을 뽑는다고 한다면 표식유전자가 조사되지 않은 경우 본 논문에서 제시한 방법에 의해 가족자료를 선발하여도 실제 정보력을 가진 가족자료가 뽑힐 가능성이 높음을 의미한다고 할 수 있다.

완전한 정보와 부분적인 표식유전자의 결측을 가진 가계자료를 함께 이용하여 분석한 결과 두 개의 표식유전자와 다섯 개의 양적형질의 10개 조합 모두에서 상위 그룹의 LOD값이 하위그룹보다 높게 나타났다. 유전율 또한 상위 그룹이 하위 그룹에 비해 높게 나타났으며 통계적으로 유의한 유전율을 보였다. 세대수가 많은 가계일수록 가족 구성원 수가 많은 경우에 높은 정보충분성을 가질 확률이 높다고 보여진다.

이 결과를 통해 본 논문에서 제시한 방법이 표식유전자가 조사되지 않은 경우에는 가계를 선별하여 추가적인 조사를 시행하여 연관성에 대한 정보를 얻어내는

기준이 될 수 있을 것이라 생각한다. 또한 표식유전자가 조사된 경우에도 조사된 가계가 의미 있는 가계인지에 대한 판단을 도울 수 있을 것이다. 하지만 이는 자료가 알려진 또는 실제 유전모형을 잘 반영하고 있다는 가정 하에서 유의한 결과를 도출할 수 있을 것이다. 더불어 표식유전자의 완전한 정보를 가진 가계와 그렇지 않은 가계를 함께 분석해도 연관성 분석에서 의미 있는 결과를 낼 수 있었지만, 정보충분성 지표의 차이가 있는 만큼 동일한 상태의 자료를 비교하는 것이 바람직하다고 생각되어 진다. 또한 본 연구의 목적이 정보력을 많이 가진 가계를 선택하는 것인 만큼 가계구조와 형질이 조사된 충분한 자료를 바탕으로 이루어질 때, 정보충분성 지표 값이 더 큰 의미를 가질 수 있을 것이다.

본 논문에서는 하나의 양적 형질과 단일 표식유전자에 국한된 결과만을 도출하였는데, 실제적으로는 유전 모형은 복합형질인 경우가 많으며 여러 유전자가 함께 작용하는 경우가 많이 있다. 때문에 여러 형질과 표식유전자를 고려한 다변량 자료에 대한 연구가 필요할 것으로 생각되어 진다. 또한 분산성분을 가장 기본적인 가법적 유전효과와 공유하는 환경적 효과, 공유하지 않은 환경적 효과로 나누어 생각하였는데, 이는 지배적 유전효과가 포함된 모형 등 다양한 성분으로 나누어 발전될 수 있다. 또한 가계자료의 횡적확장, 또는 종적확장에 따른 정보충분성 지표의 변화나 다양한 가족구성 형태에 대한 정보충분성의 연구가 필요하다고 생각되어 진다.

참 고 문 헌

박 윤주, 유전적 연관성 분석을 위한 가계도 자료의 정보 충분성에 관한 연구, 2001

박 찬미, 양적형질의 유전자 연관성 분석을 위한 개선된 헤즈만엘스톤 방법과 분산성분방법의 비교, 2003

송 기준, 양적 형질 유전자의 연관 및 관련성에 관한 동시적 분석, 2004

Almasy L. Blangero J. Multipoint quantitative linkage analysis in general pedigrees. *American Journal of Human Genetics,* 1998;62;1198-1211.

Amos C.I. Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics,* 1994;54;535-543.

Amos C.I., Dawson D.V., Elston R.C. The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. *American Journal of Human Genetics,* 1990;47;842-853.

Beaty T.H., Liang K.Y. Robust inference for variance components models in families ascertained through probands: 1. Conditioning on proband's phenotype. *Genetic Epidemiology,* 1987;4;203-210.

Blangero J., Williams J.T., Almasy L. Robust LOD score for variance component-based linkage analysis. *Genetic Epidemiology,* 2000;19(Suppl 1);S8-S14.

Boehnke M. Estimating the power of a proposed linkage study: A practical computer simulation approach. *American Journal of Human Genetics,* 1986;39;513-527.

Boehnke M., Moll P.P., Identifying pedigrees segregating at a major locus for a quantitative trait: An efficient strategy for linkage analysis. *American Journal of Human Genetics,* 1989;44;216-214.

Boehnke M., Omoto K.H., Arduino J.M. Selected pedigrees for linkage analysis of a quantitative trait: The expected Number of Informative Meioses. *American Journal of Human Genetics,* 1990;46;581-586.

Carey G., Williamsin J. Linkage analysis of quantitative traits: Increased power by using selected samples. *American Journal of Human Genetics,* 1991;49;786-796.

Dolan C.V., Boomsma D.I. A simulation study of the effects of assignment of prior identity-by-descent probabilities to unselected sib pairs, in covariance-structure modeling of a quantitative-trait locus. *American Journal of Human Genetics,* 1999;64;268-280.

Dolan C.V., Boomsma D.I. Optimal selection of sib pairs from random samples for linkage analysis of a QTL using the EDAC test. *Behavior Genetics,* 1998;28;197-206.

Eaves L., Meyer J. Locating Human quantitative trait loci: Guidelines for the selection of sibling pairs for genotyping. *Behavior Genetics,* 1994;24;443-455.

Elston R.C., Bonny G.E. Sampling considerations in the design and analysis of family studies. *Genetic Epidemiology of Coronary Heart disease,* 1984;349-371.

Elston R.C., Sobel E. Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics,* 1979;31;62-69.

Fingerlin T.E., Boehnke M., Abecasis G.R., Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. *American Journal of Human Genetic,*. 2004;74;432-433.

Fulker D.W., Cherny S.S. An Improved multipoint sib-pair analysis of quantitative traits. *Behavior Genetics,* 1996;26;527-532.

Fulker D.W., Cherny S.S., Sham P.C., Hewitt J.K. Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics,* 1999;64;259-267.

Gu C., Todorov A,m Rao D.C. Combining extremely concordant sib-pairs

provides a cost effective way to linkage analysis of quantitative traits. *Genetics Epidemiology,* 1996;13;523-533

Haydar S., Daniel E.W., Eleanor F. A survey of affected-sibship statistics for nonparametrics linkage analysis. *American Journal of Human Genetics,* 2001;69;179-190.

Kruglyak L., Lander E. S., Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics,* 1995;439-454

Ploughman L.M., Boehnke M. Estimating the power of a proposed linkage study foe a complex genetic trait. *American Journal of Human Genetics,* 1989;44;543-551.

Purcell S., Cherny S.S., Hewitt J.K., Sham P.C. Optimal sibship selection for genotypeing in quantitative trait locus linkage analysis. *Human Heredity,* 2001;52;1-13.

Rao D. C., Province Michael A. Genetic Dissection of complex traits. Academic Press.

Risch N., Zhang H. Extreme discordant sib-pairs for mapping quantitative trait loci in humans. *Science,* 1995;268;1584-1589.

Risch N., Zhang H. Mapping quantitative trait loci in humans by use of extreme concordant sib pairs: Selected sampling by parental phenotypes. *American Journal of Human Genetics,* 1996;59;951-957.

Risch N., Zhang H. Mapping quantitative trait loci with extreme discordant sib pairs: Sampling considerations. *American Journal of Human Genetics,* 1996;58;836-843.

Sham P.C., Purcell S., Cherny S.S., Abecasis G.R. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics,* 2002;71;238-253.

Sham P.C., Shao J.H., Cherny S.S., Hewitt J.K. Variance components qtl linkage analysis : Conditioning on trait values. *Genetics Epidemiology,*

2000;19(Suppl 1);S22-S28.

Olson J.M., Wijsman E.M. Linkage between quantitative trait and marker loci: Methods using all relative pairs. *Genetic Epidemiology,* 1993;10(2);87-102.

Beaty T.H., Liang K.Y. Robust inference for variance components models in families ascertained through proband: 1. Conditioning on proband's phenotype. *Genetic Epidemiology,* 1987;4(3);203-210.

Pratt S.C., Daly M.J., Kruglyak L. Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *American Journal of Human Genetics,* 2000;66(3);1153-1157.

Ekstrom C.T. Power of multipoint identity-by-descent methods to detect linkage using variance component models. *Genetic Epidemiology,* 2001;21;258-298.

Zinn-Justin A., Ziegler A., Abel L. Multipoint development of the weighted pairwise correlation(WPC) linkage method for pedigrees of arbitrary size and application to the analysis of breast cancer and alcoholism familial data. *Genetic Epidemiology,* 2001;21;41-52.

Zhao J.H.,Age of onset analysis of alcohol dependence, http://www.r-project.org (KINSHIP library)

ABSTRACT

A Study of Informativeness based on Pedigree Data for Quantitative Trait Locus Linkage Analysis

Lee, Soo Ok
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

We study pedigree selection to raise power and reduce time and expense in quantitative trait locus(QTL) linkage analysis. The method allocates a quantitative index of potential informativeness to each pedigree on the basis of observed trait scores and an assumed true QTL model. Therefore any sample of phenotypically screened pedigrees can be easily rank-ordered for genotyping. This index represents the weighted sum of χ^2 test statistics that would be obtained given the observed trait value over all possible pedigree genotypic configurations. Each configuration is weighted by the likelihood of it occurring given the assumed true genetic model. We applied our methods to the data from a cardiovascular genome center and calculated informativeness index at every pedigree. And the pedigree that have high index of informativeness have more possibility of detection in QTL linkage analysis.

Key Words : Quantitative trait locus, pedigree data, Linkage analysis, Informativeness, Variance components